BT3041: Analysis and Interpretation of Biological Data

ASSIGNMENT 1

Lecturer: Prof. Srinivasa Chakravarty

Teaching Assistant : Sandeep Nair

UTKARSH KUMAR (ME17B123)

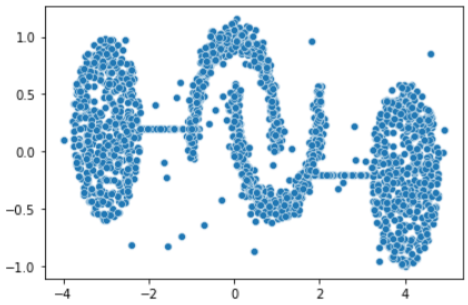
DEPARTMENT OF BIOSCIENCES

INDIAN INSTITUTE OF TECHNOLOGY MADRAS

CHENNAI, 600036

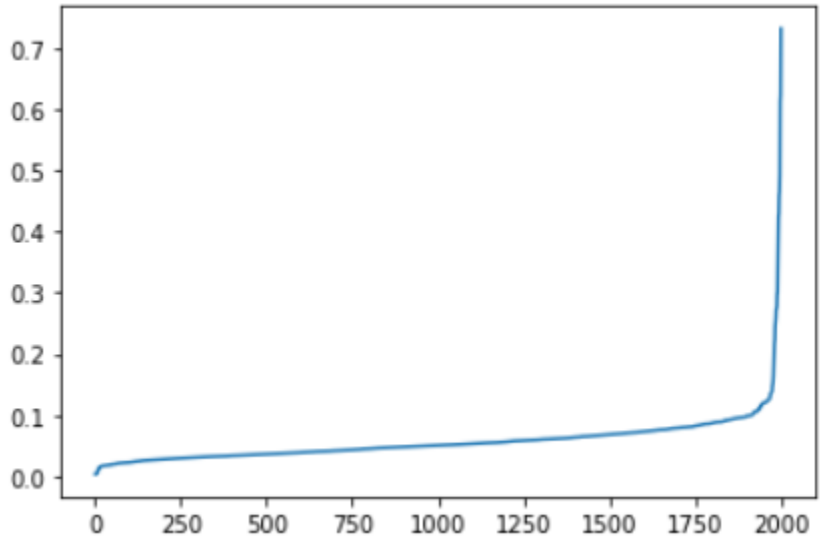
Answer 1)

The pickle file is loaded in the program and the data is plotted as follows using the seaborn function.

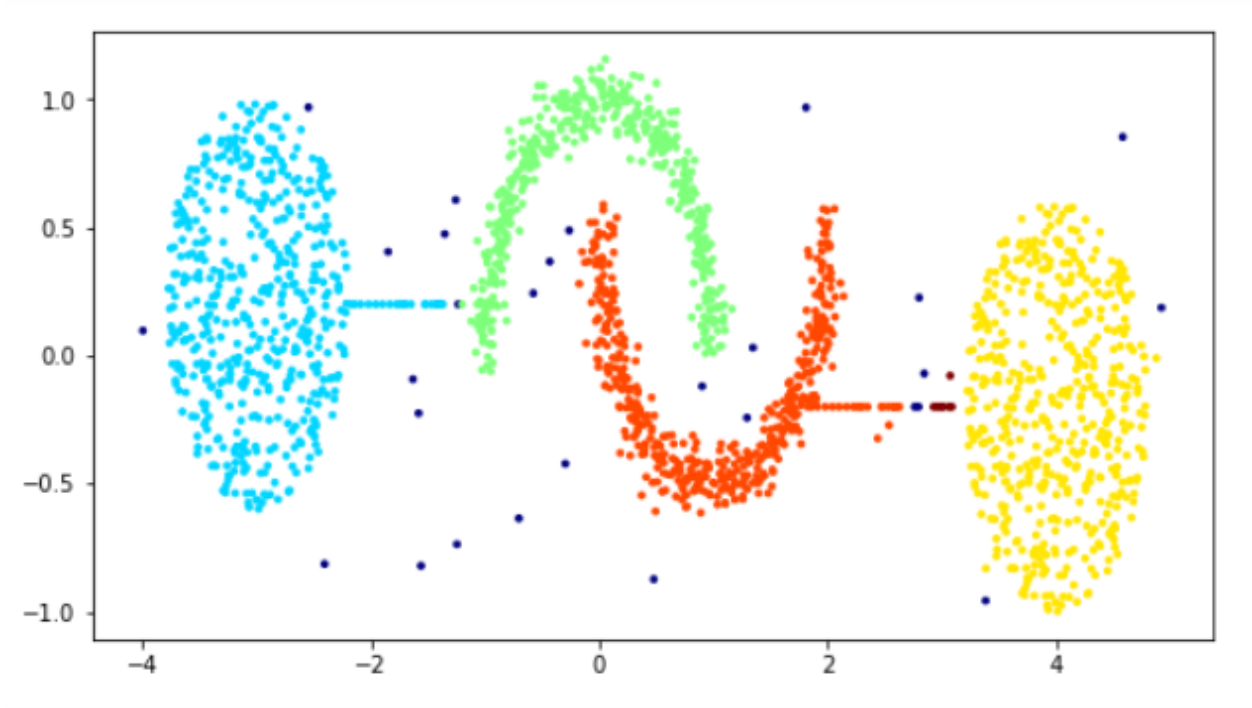


We create an algorithm called DBSCAN as given in the problem statement. Using this algorithm, we can cluster the points into separate clusters. The values of minimum points and the epsilon are not given and we need to find them.

The values of minimum points and the epsilon are found using the technique given in the book **Introduction to Data Mining (Tan, Steinbach, Kumar)**. The book describes using certain method where in most cases the value of k in the K nearest neighbours can be taken as 4. We draw a graph of the distance of K nearest neighbour from each point and then sort it, and after plotting it we obtain a graph as follows:

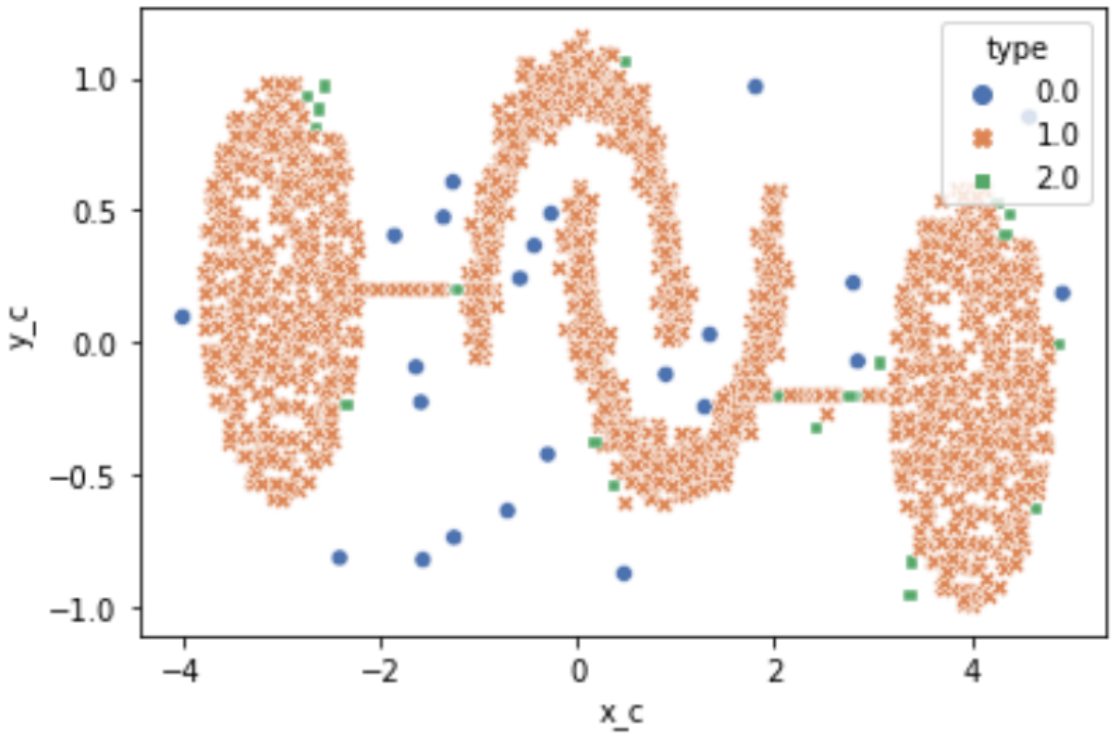


From the above graph we get to know the value of epsilon which comes out at 0.128 (Y-axis) which is the point at which the line bends abruptly. As we have already chosen the value of minimum points equal to 4, so we get the minimum points as 4 and the epsilon as 0.13 as we have rounded it off. To draw a comparison with other epsilon points in the vicinity of 0.13, we plot other clusters using epsilon starting from 0.01 to 0.15 and we can observe the different clusters that are formed in it and careful inspection tells us that epsilon with 0.13 value gives us the best clusters.



The **core, boundary and outlier** points can be found using a simple algorithm that counts the distance of each point from every other point and then uses a counter. If the counter is less than the minimum points mentioned then, we classify it as boundary point, if it is greater than minimum points then its a core point and if it isnt any of these then its an outlier.

The figure below shows the **core, boundary and outlier** points. The legend is as follows: 0 is outlier (blue), 1 is core point, 2 is boundary point.



Eps= 0.13
Minimum points= 4

Answer 2)

In this question, the .CSV file was uploaded in the code and the same algorithm of DBSCAN, which was developed from the previous question was used here.

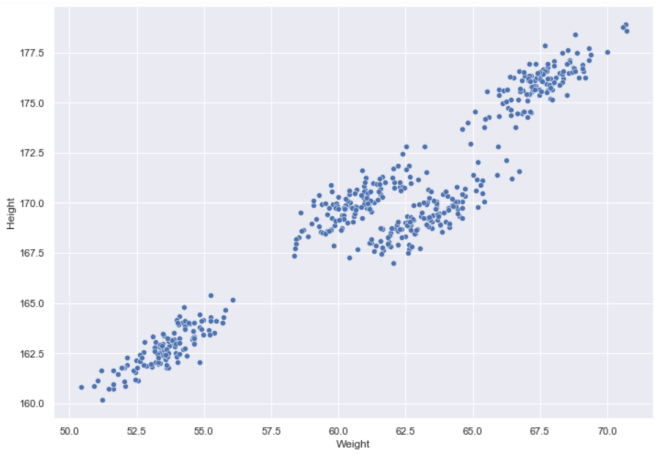


Fig: Scatter plot of given points

The weight is taken as the x axis and height as y axis. After the clustering algorithm is run, we obtain 6 clusters as shown in the following figure in different colours.

As seen in the figure above, we can find blue points scattered and others are clustered in single coloured groups.

We can find the **core, boundary and outlier** points using a simple algorithm that counts the distance of each point from every other point and then uses a counter. If the counter is less than the minimum points mentioned then, we classify it as boundary point, if it is greater than minimum points then its a core point and if it isnt any of these then its an outlier.

The figure below shows the **core, boundary and outlier** points. The legend is as follows: 0 is outlier, 1 is core point, 2 is boundary point.

An example of the core, boundary and the outliers for the above question:

BOUNDARY POINTS

```
1 X[X['Point_type']==2]
```

	Weight	Height	Point_type
4	65.431230	173.763679	2.0
6	63.341866	170.642516	2.0
8	62.633623	171.862972	2.0
12	65.163043	171.176582	2.0
15	59.176554	169.190810	2.0
...	...	...	...
491	55.223039	163.405498	2.0
492	62.311154	170.593457	2.0
494	62.200362	167.889268	2.0
496	66.423814	174.625574	2.0
498	50.433644	160.794875	2.0

224 rows × 3 columns

CORE POINTS

```
1 X[X['Point_type']==1]
```

	Weight	Height	Point_type
0	67.062924	176.086355	1.0
2	60.930863	170.284496	1.0
3	59.733843	168.691992	1.0
5	61.577160	168.091751	1.0
7	61.041643	170.096682	1.0
...	...	...	...
490	62.284684	168.673993	1.0
493	61.848894	168.260194	1.0
495	59.976983	169.679741	1.0
497	53.604698	161.919208	1.0
499	60.224392	169.689709	1.0

257 rows × 3 columns

OUTLIERS

```
1 X[X['Point_type']==0]
```

	Weight	Height	Point_type
1	68.804094	178.388669	0.0
97	69.985753	177.522585	0.0
140	67.673628	177.825299	0.0
150	56.078781	165.152797	0.0
182	64.916897	172.951507	0.0
209	66.228305	172.123237	0.0
212	66.585917	173.752915	0.0
265	65.097324	174.539183	0.0
292	51.218957	160.182164	0.0
337	62.040668	167.018251	0.0
343	63.209596	172.799210	0.0
344	65.939308	172.791283	0.0
350	58.576392	169.496276	0.0
384	54.837431	162.043844	0.0
400	68.510769	175.372059	0.0
407	59.806442	167.883527	0.0
433	65.177345	172.014466	0.0
462	65.906286	171.406661	0.0
479	55.253454	165.375646	0.0